Breast cancer risk score: a data mining approach to improve readability

Emilien Gauthier, Laurent Brisson, Philippe Lenca, Stéphane Ragusa

Abstract— **According to the World Health Organization, starting from 2010, cancer will become the leading cause of death worldwide. Prevention of major cancer localizations through a quantified assessment of risk factors is a major concern in order to decrease their impact in our society. Our objective is to test the performances of a modeling method easily readable by a physician. In this article, we follow a data mining process to build a reliable assessment tool for primary breast cancer risk. A *k*-nearest-neighbor algorithm is used to compute a risk score for different profiles from a public database. We empirically show that it is possible to achieve the same performances than logistic regressions with less parameters and a more easily readable model. The process includes the intervention of a domain expert who helps to select one of the numerous model variations by combining at best, physician expectations and performances. A risk score is made up of four parameters: *age*, *breast density*, *number of affected first degree relatives* and *prone to breast biopsy*. Detection performance measured with the area under the ROC curve is 0.637.**

I. Introduction

As cancer is becoming the leading cause of death worldwide, prevention of major types of cancer through a quantified assessment of risk is a major concern in order to decrease its impact in our society. Physicians have to inform patients about risk factors and have to detect fatal diseases as soon as possible in order to treat them as quickly as possible. Nowadays, this detection is led by prevention programs designed to target highest-risk subsets of the population. For example, women over 50 years old in France and over 40 in USA are recommended to perform a mammography every two years to detect breast cancer; mammography being the primary method for detecting early stage breast cancer which is the first cause of cancer for women [1]. As a consequence, our society could benefit from a widely used risk score in order to give more accurate counseling on how cancer is impacted by risk factors and to target smallest subset of the population with higher risks. For example, using age at first mammogram as an actionable variable, screenings programs for breast cancer could be extended: younger women with high risk profiles could be offered more frequent screenings in order to decrease death risk [2].

Even if some women may have genetic predisposition for breast cancer, environmental factors may have a larger impact on the risk according to Lichtenstein [3]. Because of this impact and due to acquisition cost and easyness-to-use constraints, we have decided to focus on environmental factors as attributes to compute a risk for women who never had breast cancer.

As pointed out by [4], "*information, dialog and more patient involvement in the decision-making process*" are key words in dealing with cancer, therefore a major challenge in the field of medical counseling is to provide physicians and radiologists with adequate tools to help them to assess breast cancer risk of their patients and to show easily how risk factors impact global risk. For many years, risk scores built upon statistical models did not reach to spread in medical counseling domain despite their performances. This may be because end-users of these tools are not oncologist nor clinician and underlying models are too complex and too difficult to use during a medical consultation. Thus, to build a new risk score tool, we need to consider the model readability and the current medical decision process. Moreover, we will have to consider the obligation to use imbalanced datasets with missing data. To the best of our knowledge, no one has been interested in analyzing, with a mining approach, data from women who never had cancer in order to create a risk score with a prevention purpose.

Showing similar cases may improve communication with the patient, therefore increase its involvement in the prevention and decision process. Because core concept of k-nearest-neighbor algorithm is to gather similar profiles using a distance computation, we use it with help of a domain expert in order to build a tool to predict breast cancer risk and measure its performances.

The paper is organized in six sections. Section II provides an overview of related works on risk models. Section III describes source data and Section IV presents our approach of the data mining process we follow. In section V, we present results, discuss them and present future works.

II. Breast cancer risk scores

A. Statistical approaches

We present studies focusing on prevention and the use of environmental factors such as reproductive and medical history. One risk prediction model emerges in the statistical field.

Based on an unstratified, unconditional logistic regression analysis, the most commonly used model was developed by

Emilien Gauthier, Laurent Brisson and Philippe Lenca are with the Institut Telecom, Telecom Bretagne, UMR CNRS 3192 Lab-STICC, Technopôle Brest Iroise CS 83818, 29238 Brest Cedex 3, France (phone: 33 2 29 00 11 75; fax: 33 2 29 00 10 30; email: {emilien.gauthier || laurent.brisson || philippe.lenca} @telecom-bretagne.eu).

Emilien Gauthier and Stéphane Ragusa are with the Statlife company, Institut Gustave Roussy, 114 rue Edouard Vaillant, 94805 Villejuif Cedex, France (phone: 33 1 42 11 51 84; fax: 33 1 42 11 40 00; email: {emilien.gauthier || stephane.ragusa} @statlife.fr).

Gail *et al* [5] using data from the *Breast Cancer Detection Demonstration Program*. Risk factor information was collected during a home interview and the analysis was based on approximately 6,000 cases and controls. Among 15 risk factors obtained through patient interviews, only 5 were chosen: age, age at menarche (first natural menstrual period), number of previous breast biopsies, age at first live birth and number of first-degree relatives with breast cancer. The model lead to the computation of a cumulative risk of breast cancer by multiplying each of the five relative risks. Then, individual risk of breast cancer is obtained by multiplication of the cumulative risk score by an adjusted population risk of breast cancer. Gail's risk score was validated on the population of United States with the *Cancer and Steroid Hormone Study* (CASH) by Costantino *et al* [6] and in Italy on the *Florence-EPIC Cohort Study* by Decarli *et al* [7].

Barlow *et al* [8] also built a risk prediction model using a logistic regression on the *Breast Cancer Surveillance Consortium* (BCSC) database (see Table I and *http://breastscreening.cancer.gov*) which contains 2.4 millions screenings mammograms and associated self-administered questionnaires (see section III). Two logistic regression risk models were constructed with 4 or 10 risk factors depending on the menopausal status. Compared to Gail's model, it gains the use of breast density and hormone therapy. As we will use the same database, it is worth highlighting that reported area under ROC curve (see performance measurement in section IV-D) was 0.631 for premenopausal women and 0.624 for postmenopausal women.

Primary goal of these studies was not readability, but rather highest risk detection performances and impact levels of each risk factors.

B. Data mining approaches and imbalanced data

Most similar data mining approaches dealt with slightly imbalanced data, mostly used to predict a cancer relapse as a result of the *Surveillance, Epidemiology and End Results* (SEER) database use. Here, we present two significant related studies involving both medical data and mining algorithm.

Endo *et al* [9] implemented common machine learning algorithms to predict survival rate of breast cancer patient. This study is based upon data of the SEER program with high rate of positive examples (18.5 %). Since this study aims at classifying examples in two classes, authors did not used ROC curve to assess performances results but accuracy, specificity and sensitivity. Logistic regression had the highest accuracy, artificial neural network showed the highest specificity and J48 decision trees model had the best sensitivity.

Jerez-Aragonés *et al* [10] built a decision support tool for the prognosis of breast cancer relapse. They used similar attributes as Gail (like age, age at menarche or first full time pregnancy, see section II-A) but also biological tumor descriptors. A method based on tree induction was conceived to select the most relevant prognosis factors. Selected attributes were used to predict relapse with an artificial neural network by computing a Bayes *a posteriori* probability in order to generate a prognosis system based on data from 1,035 patients of the oncology service of the Malaga Hospital in Spain .

TABLE I

BCSC DATABASE PUBLICLY AVAILABLE ATTRIBUTES

Full name	Short name	Description & coding
Menopausal status	menopaus	Premenopausal or postmenopausal
Age group	agegrp	10 categories from 35 to 84 years old
Breast density	density	BI-RADS breast density codes
Race	race	White, Asian/Pacific Islander, Black, Native American, Other/Mixed
Being hispanic	hispanic	Yes or no
Body mass index	bmi	4 category from 10 (underweight) to 35 and more (obese)
Age at first birth	agefirst	Before or after 30 at first live birth or nulliparous (i.e. no children)
First degree relatives	nrelbc	Number of first degree relatives with breast cancer 0, 1 or more than 2
Had breast procedure	brstproc	Prone to breast biopsy, yes or no
Last mammogramm	lastmamm	Last mammogram was negative or false positive
Surgical menopause	surgmeno	Natural or surgical menopause
Hormone therapy	hrt	Being under hormone therapy
Cancer status	cancer	Diagnosis of invasive breast cancer within one year, yes or no

Such studies show how mining approaches can be used to built classification tools on medical databases while dealing with missing data and business processes. But they do not consider problems (such as readability) encountered by patients who never had cancer nor physicians in their day to day interactions.

To build a risk score, we have to detect highest risk profiles among general population. It means we are facing highly imbalanced data with a breast cancer incidence rate lower than 1 000 new cases for 100 000 women. Dealing with such imbalanced data can be done at two levels [11], [12], [13].

At the algorithmic level, assuming all errors have a different cost is a solution to guide the data mining process [14], especially in the medical field where detecting an high risk profile is more informative than detecting a low risk profile. At the data level, sampling is another solution. A first way to rebalance data is to decrease the number of negative examples (under-sampling) [15]. And a second way of rebalancing data is to increase number of positive examples (over-sampling) [16].

III. Data source

To ensure result reproducibility, we have to choose a public database with environmental factors. The Breast Cancer Surveillance Consortium (BCSC) makes available a database that fits those major constraints. Each of the 2,392,998 lines match to a screening mammogram for a woman. This

TABLE II

MISSING DATA LEVEL BY ATTRIBUTE

Attribute	Missing data level
Body mass index	55.9 %
Age at first birth	55.5 %
Surgical menopause	52.1 %
Hormone therapy	41.0 %
Breast density	26.3 %
Last mammogramm	23.4 %
Being hispanic	20.3 %
Race	15.9 %
First degree relatives	15.2 %
Had breast procedure	10.5 %
Menopausal status	7.6 %
Age group	0 %
Cancer status	0 %

TABLE III

BREAST CANCER INCIDENCE RATE PER 100 000

Age category	SEER rate (2003-2007)	BCSC rate (1996-2002)
35-39	58.9	142.7
40-44	120.9	168.1
45-49	186.1	250.5
50-54	225.8	360.7
55-59	280.2	436.4
60-64	348.9	478.5
65-69	394.2	512.3
70-74	410.0	575.1
75-79	433.7	632.0
80-84	422.3	709.4
85+	339.2	Unavailable

publicly available database provides 12 attributes to describe the woman including cancer status.

A. *BCSC database: data collection*

Originally, the consortium was conceived to enhance understanding of breast cancer screening practices [17]. The consortium aims at establishing targets for mammography performance and a better understanding of how screenings affect patients in term of actions taken after the mammography. Domain experts from the surveillance consortium identified critical data elements for evaluating screenings performances reaching a consensus on a standard set of core data variables. Then, from 1996 to 2002, data were collected in seven centers across the United States: mammograms and their detailed analysis were collected and, at the same time, women were asked to complete a self-administrated questionnaire.

BCSC database provides personal factors (see Table I) such as factual factors (age, race, body mass index), reproductive history (age at first birth, menopausal status, hormone therapy) and medical history (number of first degree relatives with breast cancer or type of menopause). In addition, breast density was recorded when the classic Breast Imaging Reporting and Data System (BI-RADS) [18] was used by the radiologist. To ensure good quality of data, exclusion rules were set: for example, women who have undergone cosmetic breast surgery were excluded as well as women with previous breast cancer and women with no known prior mammogram.

Eventually, breast cancer cases were identified by linking cancer registries to BCSC database, i.e. for each record of the database, the class of the example is positive if the corresponding women was diagnosed with breast cancer within one year after the mammogram and completing the questionnaire and negative otherwise.

B. *BCSC database: exploratory analysis*

Among the 2,392,998 records of the database, 9,314 cases of invasive breast cancer were diagnosed in the first year of follow up. We are facing highly imbalanced data with a positive class accounting for only 0.39 % of all records.

We also observe a high level of missing data (see table II). Two main reasons explain missing data:

- Data were collected in different registries with non-standardized self-reported questionnaire: some questions were not asked and for any question, each woman had the possibility not to answer.
- Collection of some risk factors did not start at the same time. For example, height and weight were added later, explaining such a high rate of missing data for the body mass index.

Last, one has to notice that data of the BCSC are not representative of the USA breast cancer incidence rate (number of new cases during a specified time for a given population). Table III offers a comparison between the BCSC and the SEER incidence rate [19] by age categories.

Indeed, depending on data sources, the breast cancer incidence usually increase slowly from approximatively 60 to 80 years old and starts to decrease after 80 years old. But such a slower increase or decrease does not occur in the BCSC database.

IV. PROCESS TO BUILD A RISK SCORE

A. *Main objectives*

The main objective of our approach is to provide physicians with a tool to assess a cancer risk score for their patient and to promote dialog between them. As statistical models spread with difficulty in the physician community, we aim to find models with good scoring performance and good readability. In our case, we say a model has a good readability if it allows a physician to explain the risk score to his patient:

- it has to be quickly readable by a physician during a medical appointment
- and has to give access to understanding the score,

Furthermore, we have other constraints: physicians have *a priori* ideas about good attributes of a model, patients need actionable attributes to change their lifestyle, both of them want immediately usable score (i.e. very low cost of data acquisition). In addition, a generic algorithm that can be easily adapted to various pathologies is desirable.

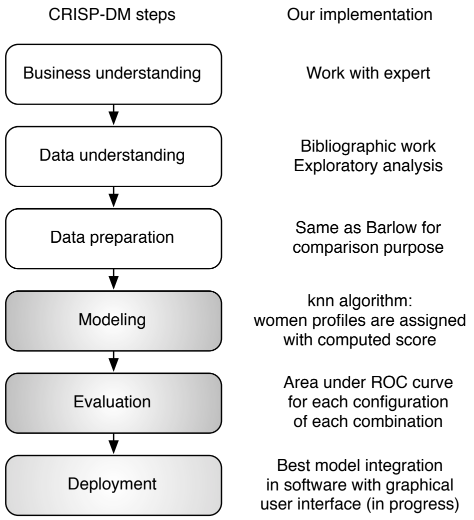

Fig. 1. General process based on CRISP-DM methodology - Gray steps identify our major contributions

B. General process

Our approach follows the CRoss Industry Standard Process for Data Mining (CRISP-DM) [20] data-mining methodology. Figure 1 shows the 6 steps of this process where gray ones identify our major contributions. Business and data understanding steps are not impacted because we want to work on the same data as [8] to be able to compare our results.

1) Business understanding: An expert with knowledge of the needs of physicians help us to prioritize our objectives (see section IV-A) and to assess the situation. We decide to focus on a scoring task (no classification or prediction).

2) Data understanding: Despite limitations described in section III, the BCSC database contains most of the known breast cancer personal factors. It is the largest database publicly available that includes breast density information.

3) Data preparation: To deal with data imbalance, we can apply rebalancing algorithms on this data but it is not the focus of the paper. We do want to minimize modification of data in order to compare our results with Barlow's. The only modification we apply is normalization. It was decided to keep the same split between training and validation set.

4) Modeling: If several data mining algorithms were considered, domain expert suggested to use a k-nearest-neighbor algorithm because it uses a concept of similarity which is easily understandable by end-users without explaining a complex formula. Moreover, such algorithm is able to deal with imbalanced data if there is enough positive examples among neighbors. We generate models and search for the best combination of attributes by performing an exhaustive search (see section IV-C) on a limited set of combinations. The reason is that the expert issued a recommendation of using a restricted number of factors to make the risk score easy to use. Obviously, for large combinations, computation time can increase sharply, but it is not a problem as models are generated offline only once by us, when a physician uses the final software, no computation is necessary.

5) Evaluation: We evaluate generated models with Receiver Operating Characteristic validation (see section IV-D) using Area Under Curve (AUC) in order to sorts models by scoring performance. Then, our expert has to choose the most useful models leveraging on the AUC performance combined with its knowledge of physician needs. ROC evaluation of every generated model is automatized in our software but we have to improve our process to formalize and support expert choice.

6) Deployment: We are currently working to incorporate selected model configuration into a computer software tool for physicians. It will come with a graphical explanation of the concept of nearest neighbor. But it will not embed the database.

C. Focus on k-nearest-neighbor implementation

To provide experts with several interesting models, k-nearest-neighbor algorithm (see [21], [22]) is used with various size of attributes combinations (from 1 to 6 attributes), several Minkowski generalized distance measure ($p = 1$ to 5) and several k values were used (see section V). Performance of each of hundreds generated combinations is tested for each values of k.

We implement the k-nearest-neighbor algorithm in two steps:

- Selection of neighborhood: for a combination of attributes (e.g. *age* and *breast density*), a score value has to be computed for each combination of values (e.g. *age=5* and *breast density=3*). To compute such score value, a neighborhood has to be defined for each values combination. To determine if a profile of the database belong to the neighborhood of a combination of values, an euclidian distance is used to compute the distance between a combination of value and every single record of the database using a normalized version of the coding values of the BCSC database. Thus, at least k of the nearest records of the database are included in the neighborhood. The neighborhood may not have always the same size because for a given group at the same distance, if k is not reached yet, all neighbors at the same distance are added to the neighborhood.
- Scoring function: the score of a combination of values, is the ratio between the number of breast cancer cases (i.e. positive examples) and the size of the neighborhood. In epidemiology, the ratio of individuals having a disease in a population is called prevalence. This ratio was chosen because it is well known by physicians, easily explainable to a patient and it is directly built on the number of patient diagnosed with breast cancer among patients with a similar profile.

To deal with missing data, we keep the same decision as Barlow, i.e. assign a high value when missing. It will prevent a record with a missing value to be integrated in the neighborhood.

TABLE IV

BEST PERFORMANCES BY COMBINATION SIZE

Size	Combinations	AUC Mean	AUC Std Deviation	AUC Median	Best combination (See Table I)	AUC
1	12	0.536	0.030	0.529	agegrp	0.614
2	66	0.563	0.031	0.553	agegrp+density	0.635
3	220	0.581	0.029	0.601	agegrp+density+brstproc	0.641
4	495	0.593	0.026	0.597	agegrp+density+brstproc+lastmamm	0.642
5	792	0.602	0.023	0.586	agegrp+density+brstproc+lastmamm+menopaus	0.642
6	924	0.607	0.019	0.603	agegrp+density+brstproc+lastmamm+hrt+nrelbc	0.637

D. Focus on ROC evaluation

The Receiver Operating Characteristic (ROC) [23] is used to measure performance due to the continuous nature of our classifier: performance has to depict how positive instances are assigned with higher scores than negative ones. The ROC curve allows to measure detection performances using a moving threshold to classify examples of the validation set. Moreover, it allows direct comparison with Barlow's results and epidemiological-based scores in general.

Negative examples labeled as positive by the algorithm are called a false positives whereas positive examples labeled as positives are called true positives. The ROC curve is plotted with the false positive rate on the X axis and the true positive rate on the Y axis [24], both rates being calculated for a given threshold. It can be summarized in one number: the Area Under the ROC Curve (AUC). The area being a portion of the unit square, its value is in then [0,1] interval. The best classifier will have an AUC of 1.0 (i.e. all positive examples are assigned with higher score than negative ones) whereas an AUC of 0.5 is equivalent to random score assignment.

Each k value of each combination of attributes is assigned with a ROC curve and the corresponding AUC in order to help the expert to choose the best model.

V. EXPERIMENTAL RESULTS

A. Scoring performances

An experiment set was designed to test how the k-nearest-neighbor algorithm perform on the BCSC data. As one of our constraint is to build a readable risk score (see section IV-A), we select all combinations with a size s of 1 to 6 attributes among $n = 12$ available attributes, meaning we have $\sum_{s=1}^{6} \frac{n!}{s!(n-s)!} = 2,509$ combinations to test. A first way of assessing results of these combinations is to look at the best combinations by size (see Table IV). These results are obtained in an euclidian space using a 2-norm euclidian distance as they are not significantly better, when improved, using another p-norm measures.

Among one attribute combinations, *agegrp* is by far the best factor to score breast cancer risk in the BCSC database with an AUC of 0.614, while the next best attribute (not shown), *menopaus* for menopausal status, performs only at 0.563. This result confirms expert knowledge since it's widely known that age is a major breast cancer risk factor.

For combinations size from 1 to 3 attributes, mean, median and best AUC rise, whereas for sizes of 4 and 5 attributes, maximal performances level off around 0.64 with a slight decrease with 6 attributes for best combinations. It is interesting to obtain the best results using less possible attributes to improve model readability. Furthermore, our 3 attributes *agegrp, density, brstproc* combination has an AUC of 0.641 while in Barlow's results (see section II-A), at least 4 attributes are needed to achieve an AUC of 0.631 on a subset of data that includes only premenopausal women only.

TABLE V

TOP 15 PERFORMANCE RESULTS BEFORE AND AFTER EXPERT ADVICE

A. **Best combinations before expert advice**	AUC
agegrp, lastmamm, density, brstproc	**0.642**
menopaus, agegrp, lastmamm, density, brstproc	0.642
agegrp, density, brstproc	0.641
menopaus, agegrp, density, brstproc	0.641
bmi, agegrp, density, brstproc	0.640
bmi, agegrp, lastmamm, density, brstproc	0.640
agegrp, hispanic, density, brstproc	0.640
agegrp, density, brstproc, agefirst	0.639
agegrp, hispanic, lastmamm, density, brstproc	0.639
bmi, agegrp, density, brstproc, race	0.638
menopaus, agegrp, hispanic, density, brstproc	0.638
hrt, agegrp, lastmamm, density, brstproc	0.638
agegrp, density, brstproc, race	0.638
agegrp, surgmeno, lastmamm, density, brstproc	0.638
agegrp, lastmamm, density, brstproc, race	0.638

B. **Best combinations after expert advice**	AUC
agegrp, density, brstproc	0.641
menopaus, agegrp, density, brstproc	0.641
bmi, agegrp, density, brstproc	0.640
agegrp, hispanic, density, brstproc	0.640
agegrp, density, brstproc, agefirst	0.639
bmi, agegrp, density, brstproc, race	0.638
menopaus, agegrp, hispanic, density, brstproc	0.638
agegrp, density, brstproc, race	0.638
menopaus, agegrp, surgmeno, density, brstproc	0.638
agegrp, hispanic, density, brstproc, agefirst	0.638
bmi, agegrp, hispanic, density, brstproc	0.638
menopaus, agegrp, density, brstproc, agefirst	0.638
bmi, agegrp, density, brstproc, agefirst	0.637
menopaus, hrt, agegrp, density, brstproc	0.637
agegrp, density, brstproc, nrelbc	**0.637**

A first list of all possible combinations (from 1 to 6 attributes), is produced and sorted by performances (see Table V-A). We observe that with an AUC of 0.642, the *agegrp, density, brstproc, lastmamm* combination perform better than the two specialized regression models obtained on pre- and postmenopausal women by [8].

B. Use of expert knowledge

As stated in section IV-A, besides scoring performances, our main objectives also include readability and integration of *a priori* ideas from physicians. This step of the process involves contribution from a domain expert (see section IV-B).

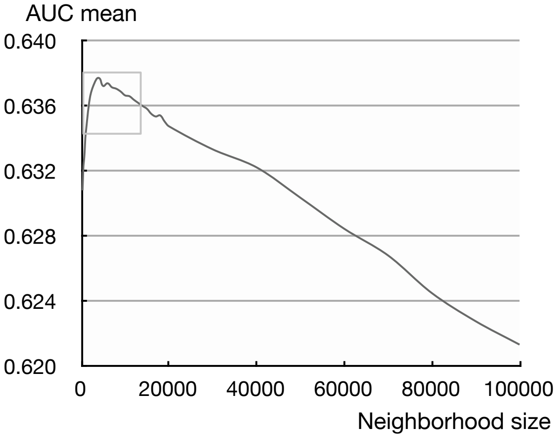

Fig. 2. Performances of top 15 combinations from Table V-B

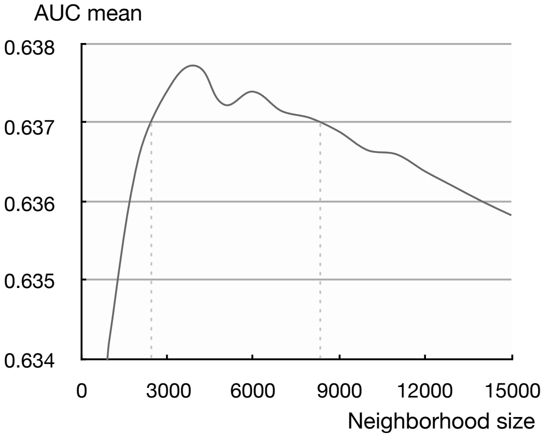

Fig. 3. Zoom on performances of top 15 combinations from Table V-B

From our domain expert point of view, with Table V-A in hand, it appears that the result of the last mammogram is a costly piece of information to obtain from women during a counseling appointment with a physician compared to performance improvement. Domain expert chooses to reduce his choices list to available combinations without *lastmamm*. Top 15 performances measures without *lastmamm* attribute are shown in Table V-B.

Based on his domain knowledge, the expert highlights that the number of first degree relatives affected by breast cancer (*nrelbc*) is widely recognized as an important factor in breast cancer risk whereas other risk factor, like the body mass index (*bmi*), are not that important compared to others. According to this expert, a good candidate for our risk score would be the *agegrp, density, brstproc, nrelbc* combination with an AUC of 0.637, which is a good performance compared with best performances of Barlow's logistic regression model (AUC of 0.624 to 0.631 depending on menopausal status). This combination uses relevant attributes for physicians according to our expert and performance loss, from 0.642 to 0.637, is acceptable.

C. Stability

In order to run a k-nearest-neighbor algorithm, the size of neighborhood has to be set. Since only k closest neighbors are used to compute the ratio healthy vs. diseased, risk score value depends on k value. If the neighborhood is too small, few breast cancer cases are included and if the neighborhood is too large, patient profiles are too different: in both cases the risk score is not reliable. For each of the 2,509 combinations of attributes, we tested the scoring function with 40 values of k from 100 to 100 000.

Using, as an example, the top 15 combinations from Table V-B, we plotted the evolution of the performance (using the AUC mean) depending on the size of the neighborhood (see Fig. 2). With an undersized neighborhood, performances are low but then, as k increases, performances increase with a maximum of 0.638. From 2,500 to 8,400 neighbors (see Fig. 3), performances are always higher than 0.637 meaning that the algorithm is relatively stable depending on k and ultimately on the number of positive examples in the neighborhood. Eventually, as k increases, performances decrease because using a larger neighborhood leads to compute a ratio with increasingly dissimilar profiles and poor targeting.

It means that performance of the combination is not obtained with a local maximum for a single value of k. It rather depicts overall prediction ability of a combination independently of the value of k as long as the size of the neighborhood is large enough to be statistically reliable (according to the law of large numbers) and stringent enough to eliminate too dissimilar profiles.

D. Discussion

As statistical risk scores do not spread in the medical community, we think there is a possibility to improve risk scores to offer both readability in its elaboration and possibility for experts to integrate their knowledge (regarding end users expectations and the disease itself) in the process. A standard methodology called *CRISP-DM* was followed in the process of building such a risk score. The database from the BCSC was selected because a regression-based score was already built upon it and because the database itself was publicly available. We chose to run extensive test with a k-nearest-neighbor algorithm to score profiles with different combinations of attributes. Every combinations with 1 to 6 attributes were tested, each for several values of k neighbors. Thus, we were able to allow experts to establish rules to keep or reject combinations by weighting between performance versus attributes usefulness and risk factors expected by physicians.

Nevertheless, our study has some limitations. First, even if we selected one of the few databases large enough to be representative of the targeted population, findings from database of volunteers require cautious extrapolation to general population. Second, if the concept of similarity used in the algorithm is easy to understand for everyone, performances may be limited due to imbalanced data and the constraint of not modifying data used in this paper in order to be able to compare results. However, options are available to improve steps of the process. Better performances may be obtained using another algorithm, potentially with balance of data in the data preparation step, or by combining k-nearest-neighbor with another algorithm [25]. Use of expert knowledge could be improved by selecting models which are provided to the expert to avoid complications due to the size of the list of combinations. Performances could also be improved by integrating domain knowledge deeper in the algorithm: for example, introduction of relative risk as a weight in the distance computation may help to deal with the different level of influence of each risk factors.

Since k-nearest-neighbor algorithm gives good results, we think it would be useful to test this process on another database that include continuous attributes that were not discretized. For example age or breast density are one of the most predictive attributes and more specific data should improve performances. Higher risk profiles should be more accurately targeted leading to increased performances.

VI. Conclusion

On a medical dataset, we obtain good results on readability on the modeling method with a k-nearest-neighbor algorithm easy to understand for physicians and patients. In addition, the score is very easy to use for end-users with only four attributes needed. We also allow the expert to choose a combination that has not necessarily the best detection performance, but show qualities like physician acceptance and inclusion of most performant attributes recognized by the community.

Our approach is innovative and successful because we have shown that it is possible to build a simple and readable risk score model for primary breast cancer prevention that performs as good as widely used logistical models.

References

[1] "World Cancer Report," p. 512, 2008. [Online]. Available: http://www.iarc.fr/en/publications/pdfs-online/wcr/index.php

[2] F. C. Teams, "Mammographic surveillance in women younger than 50 years who have a family history of breast cancer: tumour characteristics and projected effect on mortality in the prospective, single-arm, fh01 study," *The Lancet Oncology*, vol. 11, no. 12, pp. 1127–1134, 12 2010.

[3] P. Lichtenstein, N. V. Holm, P. K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, and K. Hemminki, "Environmental and heritable factors in the causation of cancer, analyses of cohorts of twins from sweden, denmark, and finland," *New England Journal of Medicine*, vol. 343, no. 2, pp. 78–85, 07 2000.

[4] P. Testard-Vaillant, "The war on cancer," *CNRS international magazine*, vol. 17, pp. 18–21, 2010.

[5] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, and J. J. Mulvihill, "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually," *J. Natl. Cancer Inst.*, vol. 81, no. 24, pp. 1879–1886, 1989.

[6] J. Costantino, M. Gail, D. Pee, S. Anderson, C. Redmond, J. Benichou, and H. Wieand, "Validation studies for models projecting the risk of invasive and total breast cancer incidence." *J Natl Cancer Inst*, vol. 91, no. 18, pp. 1541–8, 1999.

[7] A. Decarli, S. Calza, G. Masala, C. Specchia, D. Palli, and M. H. Gail, "Gail model for prediction of absolute risk of invasive breast cancer: Independent evaluation in the florence-european prospective investigation into cancer and nutrition cohort," *J. Natl. Cancer Inst.*, vol. 98, no. 23, pp. 1686–1693, 2006.

[8] W. E. Barlow, E. White, R. Ballard-Barbash, P. M. Vacek, L. Titus-Ernstoff, P. A. Carney, J. A. Tice, D. S. M. Buist, B. M. Geller, R. Rosenberg, B. C. Yankaskas, and K. Kerlikowske, "Prospective breast cancer risk prediction model for women undergoing screening mammography," *J. Natl. Cancer Inst.*, vol. 98, no. 17, pp. 1204–1214, 2006.

[9] A. Endo, T. Shibata, and H. Tanaka, "Comparison of seven algorithms to predict breast cancer survival," *Biomedical Soft Computing and Human Sciences*, vol. 13 2, pp. 11–16, 2008.

[10] J. M. Jerez-Aragonés, J. A. Gómez-Ruiz, G. Ramos-Jiménez, J. Muñoz-Pérez, and A.-C. E., "A combined neural network and decision trees model for prognosis of breast cancer relapse," *Artificial Intelligence in Medicine*, vol. 27, pp. 45–63(19), jan 2003.

[11] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[12] S. Visa and A. Ralescu, "Issues in mining imbalanced data sets - a review paper," in *Sixteen Midwest Artificial Intelligence and Cognitive Science Conference*, 2005, pp. 67–73.

[13] G. M. Weiss and F. Provost, "Learning when training data are costly: The effect of class distribution on tree induction," *J. of Art. Int. Research*, vol. 19, pp. 315–354, 2003.

[14] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *Fifth International Conference on Knowledge Discovery and Data Mining*. ACM Press, 1999, pp. 155–164.

[15] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory under-sampling for class-imbalance learning," in *International Conference on Data Mining*, 2006, pp. 965–969.

[16] A. Liu, J. Ghosh, and C. Martin, "Generative oversampling for mining imbalanced datasets," in *International Conference on Data Mining*, 2007, pp. 66–72.

[17] R. Ballard-Barbash, S. Taplin, B. Yankaskas, V. Ernster, R. Rosenberg, P. Carney, W. Barlow, B. Geller, K. Kerlikowske, B. Edwards, C. Lynch, N. Urban, C. Chrvala, C. Key, S. Poplack, J. Worden, and L. Kessler, "Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database," *Am. J. Roentgenol.*, vol. 169, no. 4, pp. 1001–1008, 1997.

[18] V. Reston, Ed., *Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas)*. American College of Radiology, 2003.

[19] S. F. Altekruse, C. L. Kosary, M. Krapcho, N. Neyman, R. Aminou, W. Waldron, J. Ruhl, N. Howlader, Z. Tatalovich, H. Cho, A. Mariotto, M. Eisner, D. R. Lewis, and B. K. Edwards. (2010, August) SEER Cancer Statistics Review, 1975-2007. http://seer.cancer.gov/csr/1975_2007/.

[20] P. Chapman, J. Clinton, R. Kerber, and T. Khabaza, "Crisp-dm 1.0 step-by-step data mining guide," The CRISP-DM Consortium, Tech. Rep., 2000.

[21] E. Fix and J. Hodges, "Discriminatory analysis, non-parametric discrimination: consistency properties," USAF Scholl of aviation and medicine, Randolph Field, Tech. Rep., 1951.

[22] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[23] J. P. Egan, *Signal detection theory and ROC analysis*, ser. Series in Cognition and Perception. Academic Press, 1975.

[24] T. Fawcett, "An introduction to ROC analysis," *Pattern Recogn. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[25] N.-K. Pham, T.-N. Do, P. Lenca, and S. Lallich, "Using local node information in decision trees: Coupling a local labeling rule with an off-centered entropy," in *International Conference on Data Mining*, 2008, pp. 117–123.